A scientific methodology for researching CALL interaction data: Multimodal LEarning and TEaching Corpora

Thierry Chanier, Ciara Wigham

Chanier, Thierry and Wigham, Ciara R. (2016). "A scientific methodology for researching CALL interaction data: Multimodal LEarning and TEaching Corpora". In Caws, Catherine, and Hamel, Marie-Josée (eds), *Language-Learner Computer Interactions: Theory, methodology and CALL applications*. John Benjamins. Pp. 215-240, DOI: 10.1075/lsse.2.10cha

A scientific methodology for researching CALL interaction data: Multimodal LEarning and TEaching Corpora

Thierry Chanier, Université Blaise Pascal Clermont-Ferrand, France
Ciara R. Wigham, Université Lumière Lyon 2, France

Abstract

This chapter gives an overview of one possible staged methodology for structuring LCI data by presenting a new scientific object, LEarning and TEaching Corpora (LETEC). Firstly, the chapter clarifies the notion of *corpora*, used in so many different ways in language studies, and underlines how corpora differ from raw language data. Secondly, using examples taken from actual online learning situations, the chapter illustrates the methodology that is used to collect, transform and organize data from online learning situations in order to make them sharable through open-access repositories. The ethics and rights for releasing a corpus as OpenData are discussed. Thirdly, the authors suggest how the transcription of interactions may become more systematic, and what benefits may be expected from analysis tools, before opening the CALL research perspective applied to LCI towards its applications to teacher-training in Computer-Mediated Communication (CMC), and the common interests the CALL field shares with researchers in the field of Corpus Linguistics working on CMC.

Keywords: LEarning and TEaching Corpora (LETEC), staged methodology, multimodal transcription, OpenData

Introduction

In many disciplines, research is based on the availability of large research data sets, built collaboratively from the work of many different research teams. Data are shared and form the basis for new analyses, or counter-analyses. To meet this demand for data, other researchers develop tools for the research cycle (tools for capturing and analysingdata). When studying Learner Computer Interactions (LCI), researchers are concerned by the extent of data collection and by the description of the context in which data were collected. Studying online learning, in order to understand this specific type of situated human learning and/or to evaluate pedagogical scenarios or technological environments, requires accessibility to interaction data collected from the learning situation.

The intention of this chapter is to give an overview of one possible staged methodology for structuring LCI data. It presents a new scientific object, the *LEarning & TEaching Corpora* (LETEC). After having clarified the notion of corpora, used in so many different ways in language studies, the methodology used to collect, transform and organize data in order to make them sharable through open-access repositories is described. We suggest how the transcription of interactions may become more systematic, and what benefits may be expected from analysis tools before opening the CALL research perspective applied to LCI towards its applications to teacher-training in Computer-Mediated Communication (CMC), and the common interests we share with researchers in the field of corpus linguistics working on CMC.

Differentiating raw language data and corpora

Corpora in Linguistics

In many areas of general linguistics or even applied linguistics, building and using a corpus is a tradition. A first definition offered by Biber, Conrad and Reppen (1998), following the seminal work of Sinclair (1991) (see O'Keefe et al. [2007] for full references), could be as follows: a corpus is a principled collection of texts, written or spoken, available for qualitative or quantitative analysis. The word corpus, however, may be indistinctly used by a graduate student to refer to her/his compilation of a set of language examples or a set of texts, or by a researcher in corpus linguistics. A similar confusion exists inthe Humanities around the word database. Any set of data included in a spreadsheet, or even database software, is often labelled a database, while the second indispensable component of a database, i.e., its conceptual model or semantic level, is ignored. This model, also developed by the data compiler, is often considered as the most valuable component because, firstly, it brings data up to the level at which it may be considered as information and, secondly, because it allows queries and computations to be executed on the basic level of data.

Coming back to language issues, Bernard Laks, a scholar in speech corpora, often underlines the amount of time (over thirty years) it took for linguists to shift from the exemplum paradigm to the datum paradigm (Laks, 2010). At the end of the fifties, a number of linguists, influenced by Chomsky, rejected the idea of working on corpora (perceived as "limited" in nature) and based their analyses only on sets of language examples, which sometimes were even invented in order to include what they considered as interesting phenomena. Today, many linguists consider that language should be studied in contexts of real usage and, consequently, that corpora are the way to capture language usage.

The nature of corpora and the methodologies for building them have largely evolved from the seminal work of Kucera and Francis(1964) who designed the Brown Corpus as a reference corpus for American English. For example, the DWDS (*DigitalesWörterbuch der DeutschenSprache*, 2013) corpus of modern German contains billions of tokens/words; teams of linguists, who have patiently chosen the various genres that reflect the way German is currently used (including Internet genres), have solved issues

concerning rights access and collected the data. Raw data are never compiled as such, but rather transferred into standard formats, based on the *eXtensibleMarkup Language* (XML). Researchers developed XML schemas, which play a similar role to the conceptual model of databases. XML is used on top of the texts, sentences and words to add annotations.

Corpora in CALL

The language-teaching domain is also directly concerned with corpora. Launched in the nineties, conferences including TALC (*Teaching And Language Corpora*) have become popular among applied linguists, and some language teachers are interested in the idea of using different kinds of language corpora in their teaching (O'Keefe et al., 2007). As an example, if German academic writing is considered, linguists may study this type of language for specific purposes (LSP) before updating pedagogical handbooks with language structures that are actually used, or teachers may use the same LSP corpora with learners of German. The latter situation is often referred to as Data-Driven Learning (DDL) (Boulton, 2011), and special interest groups within the CALL community have developed in this area, as well as dedicated journal issues.

Whereas the previous corpora all captured language used in formal or informal situations only by native speakers, a team of linguists gathered in Belgium around Sylviane Granger to launch a new type of corpora, namely Learner Corpora. Productions (mainly academic essays) of learners of English as a second language were collected (Granger, 2004). Here again, the team did not confuse the concept of a corpus with a simple set of essays in electronic formats. They developed a framework for learner corpus research where data were collected, structured and, from 2009 onwards, annotated in the same way. They included productions of learners with different mother tongues to allow interlanguage comparisons.

The corpus paradigm

Taking into consideration the aforementioned lengthy experiences coming from corpus linguistics (whether general or applied linguistics), as well as international recommendations for the management of research data in all scientific disciplines, the corpus paradigm can be developed as follows:

- *Systematic Data Collection*. Even when an individual researcher has a specific research question in mind, let us say, a specific kind of interaction s/he wishes to consider, the whole data set, including interactions, productions, logfiles (data related to what is called learning analytics) should be collected. It is a prerequisite to allow other researchers to reuse the corpus. It also relates to quality criteria. Often a researcher selects a subset of data from the whole data set in order to analyse it and publish an article. Quality in the research procedure implies that the researcher is able to explain the extent to which a selected subset of data does not correspond to a simple disconnected episode, but really reflects what happened during the online course.

- *Detailed Data Description*. The context of learning situations encompasses many facets, as detailed later in this section. In regards to language corpora,in general, the detailed descriptions are often referred to as metadata. In the metadata, the researcher not only gives a corpus title, date, list of credits, but also explains how the data have been collected, edited and organized. Sociolinguistic information about the participants is detailed. As an example, let us consider a SMS corpus. Metadata will explain how messages have been collected on the phone network(s) and anonymised. They will document participants who sent the messages, the structure of the messages assembled in the body/text of the corpus, whether the date of a message corresponds to its date of posting or of collection, and the way in which IDs have been attributed to participants to guarantee that messages sent by the same person can be linked, etc. This information is essential if a researcher wishes to carry out a discourse-analysis study.

- *Data Conversion*. Time spent on data collection and description will be valued during the analysis phase. It is now generally considered as a multiple-step process, where output of a first analysis tool will become input for a second tool. Young researchers working on language-related data, whether oral, textual or multimodal (optionally, incorporating non-/co-verbal data), will often have to manage this analysis flow before the publication, for example, of her/his thesis. This has two main implications: (a) the use of analysis tools that accept open formats for data input and that do not produce output in proprietary formats, and (b) conversion, organization and structuring of the collected data into standard formats. Besides open-access formats for images, audio or video files, the format for textual data is now based on XML, not simply a basic XML level, but levels of higher standards that allow annotations and multi-level analyses, as detailed further on.

- *Data Release and Distribution*. As previously explained, a language corpus and its related analysis can only become part of the scientific research cycle if it can be freely accessed and when this access is guaranteed as permanent. Although solutions and access to procedures that guarantee this openness are well known, availableand fairly simple, the current situation is blurred by the misuse of the term OpenData (see a relevant definition in Open Knowledge, 2013, as well as Chanier, 2013). If a researcher tries to access language corpora which pretend to be open access, s/he may discover free access to only a limited part of the corpus, or that the corpus cannot be downloaded, or, when it is a speech corpus, s/he may only have access to the transcripts but not the accompanying audio files, etc. Under such circumstances, research on the corresponding data is impossible. However, there currently exist more frustrating situations—for example, when a researcher adds an extra level of annotation and wants to publish this, but suddenly realizes that s/he is not allowed to because the license attributed by the original collectors of the corpus forbids any derivative work. Securing open access intertwines several steps of a corpus' lifecycle. Before data are collected, the researcher will consider the question of ethics and rights related to participants and their productions, choose the license under which to release the future corpus, and choose in which repository the corpus will be deposited for archiving, for example, at the European level, DARIAH (2013).

Clarifying some terms.

Before considering corpora specific to LCI, definitions of terms used in many different ways across the field of linguistics, as well as in other disciplines, need to be elicited (see Chanier et al., 2014).

Firstly, the word text is interpreted here in its broad sense relating to its multimodal nature, with respect to Baldry and Thibault (2006) who defined texts as "meaning-making events whose functions are defined in particular social contexts" (p. 4), and Halliday (1989) who declared that "any instance of living language that is playing a role some part in a context of situation, we shall call it a text. It may be either spoken or written, or indeed in any other medium of expression that we like to think of"(p. 10). Simply stated, learners compose a text when they produce utterances, for example, in an audio chat.

Secondly, an online environment may be synchronous or asynchronous, mono- or multimodal. Modes (text, oral, icon, image, gesture, etc.) are semiotic resources that support the simultaneous genesis of discourse and interaction. Attached to this meaning of mode oriented towards communication, we use the term modality as a specific way of realizing communication as per the Human Computer Interaction field (Bellik&Teil, 1992). Within an environment, one mode may correspond to one modality, with its own grammar constraining interactions. For example, the icon modality within an audio graphic environment is composed of a finite set of icons (raise hand, clap hand, is talking, momentarily absent, etc.). In contrast, one mode may correspond to several modalities: Text chat has a specific textual modality that is different from the modality of a collective word processor, although both are based on the same textual mode. Consequently, an interaction may be multimodal because several modes are used and/or several modalities.

After having considered criteria for general types of language corpora, the next section presents criteria specific to LCI illustrated by the LETEC approach.

An illustration of the staged methodology for building LETEC

The LETEC approach to data collection, structuring and analysis comprises successive phases (Figure 1). It has been developed from 2006 onwards by the Mulce project (Reffay et al., 2012). Using a case-study approach, this section describes these phases in turn, referring to the example of the online English for Specific Purposes course, Copéas, and its associated LETEC (see Chanier et al., 2009). This ten-week intensive course ran in 2005 and formed part of a master's programme in Distance Education in France. The course's aims were for students to be able to think critically about using the web for learning and to practise their oral and written English skills online. Each week, the students participated in online tutored discussions in the online platform Lyceum.

Lyceum is an audio-graphic conferencing environment that included communication modalities (audio chat, text chat, iconic system) and shared editing modalities (whiteboard, concept map, shared word processor). For the reasons already given, it was a multimodal environment,as shown inFigure 10.1, and explained in Ciekanski and

Chanier (2008). Lyceum no longer exists. However, thanks to the availability of LETEC data, the environment's features, as well as how participants used it to work and communicate, can be studied and compared to other environments.

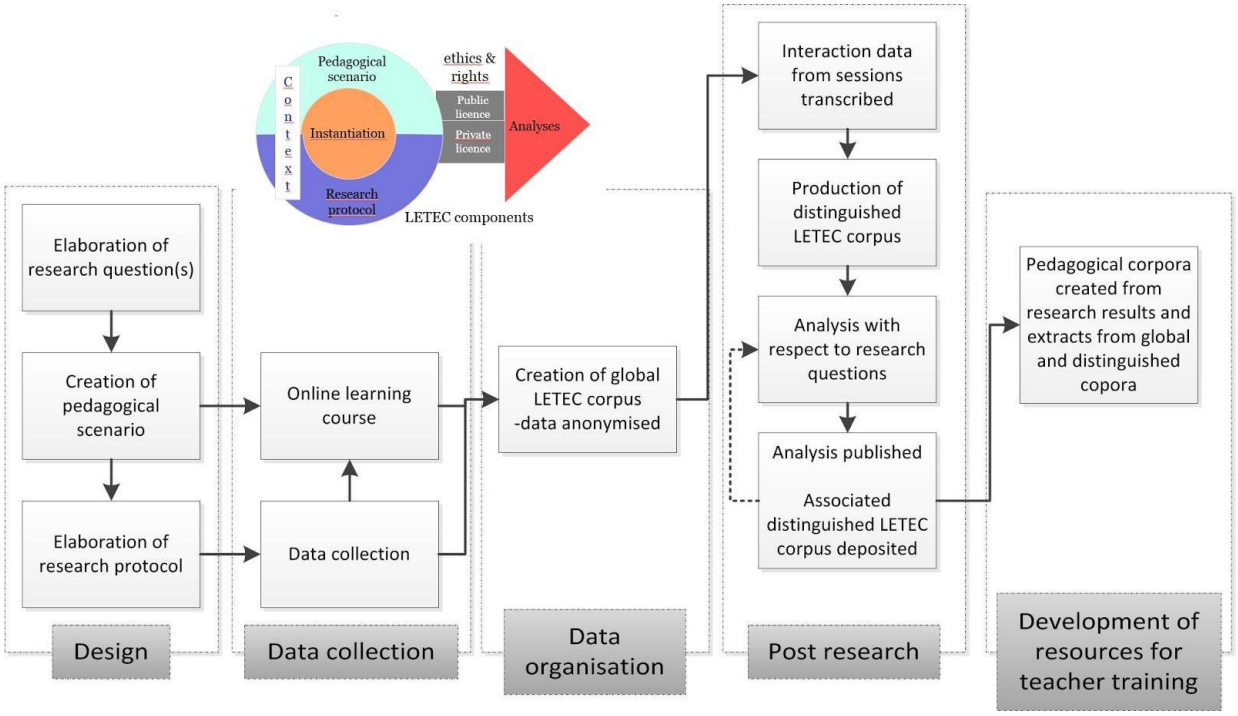

Figure 10.1.Successive phases of a LETEC approach to an online learning situation. LETEC components are illustrated in the top-left hand schema.

Design: Pedagogical scenario and research protocol

The first stage of a LETEC methodological approach is to determine the research focus. That is to say the type of phenomenon concerned and the aspect that is of interest. At this stage, it is important to imagine the possible end product that is initially intended. The Open University (2001)has examined a range of general purposes for conducting educational research: to describe, explain, predict, evaluate, prescribe and theorize (p. 30). Identifying a clear research purpose will influence how the research questions are formulated, the type of data to be investigated and how the researcher can select these. Although the research focus will be determined at the beginning of the research process, it is important to note that research questions may not be formulated until later on, or, if formulated during the design phase, they may be modified in between the LETEC design stage and the post-research analysis stage and will most likely become more focused.

In parallel to determining the research focus and specifying the research questions, the online learning context in which it will be examined needs to be elaborated. The design of an online learning situation requires the creation of a pedagogical scenario. This describes (a) the whole online environment (such as a Learning Management System

[LMS], a videoconferencing system and their different subcomponents); (b) the various roles the participants (teachers, learners, experts, such as natives, etc.) will undertake during the course; (c) each course activity and the role of each participant during this (e.g., one learner may act as a group animator/tutor) and the component of the online environment the activity is linked to; (d) how activities are sequenced (the workflow); (e) the resources that will be used and produced; and (f) the instructions that govern the learning activities. To avoid confusion between the role of the participants who are involved in supporting the learners and the learning tasks, the pedagogical scenario may consist of a learning scenario and a tutoring/supervision scenario, the latter detailing how different participants will aid learning and how teachers/tutors will intervene during the course in supervisory actions. Put simply, the pedagogical scenario will answer the question of *who does what, when, with what tools and for what results* (see IMS-Learning Design in IMS-Learning, 2004).

If the online learning situation is to be the focus of a research study, it is also necessary to elaborate a research protocol. This will take into account the variables to investigate, the participants in the study, human subject ethical protections, the methods and procedures to be used for data collection and any reliability or validity of collection methods. In relation to the pedagogical scenario, the research protocol details moments at which activities uniquely related to research will be accomplished (e.g., consent form distribution, pre- and post-course questionnaires, post-course interviews). If observation is to occur, the role of the researcher(s) will also be determined.

The pedagogical scenario and the research protocol could be described as a simple textand assembled with all the documents (pedagogical guidelines, instructions given to teachers, learners, questionnaires forms, etc.); however, this description has to be detailed. It represents more than the usual context of interactions. Research in CALL studies the influence of the learning situations on the interactions and their outcomes. Hence, scientific investigation can commence only if the learning context is explained in a way that a researcher who did not participate in the course could understand the situation. This is why it is recommended to use standard[1]formats for describing these elements, particularly formats that allow visual presentations of the pedagogical scenario, the research protocol and that allow links to resources (IMS-Learning, 2004).

Data collection

After planning the online learning situation and the research design, the next phase is to systematically gather the data. Data collection focuses on acquiring information, in an ethical manner, to attempt to answer the research questions elaborated during phase one of the LETEC approach and with reference to the research protocol established.

---

[1]The word *standard* is frequently used in this chapter to refer to formats which are shared among academic communities to describe different levels of information within corpora. When large sets of communities agree upon a standard, it may become an international *norm* (such as those used by ISO – International Standard Organization). Useful standards generally need to be open (not attached to proprietary formats) and accepted by a wide range of software analysis tools (asset often called *interoperability*).

This phase has to be carefully planned beforehand. Earlier on, we mentioned decisions that have to be made before collection and which may influence other phases: interaction data may be difficult to extract from some environments but easier from others that have the same affordances; data formats generated by the learning environments or from other recording devices (audio recorder, screen capture software, etc.) should be easy and not too time-consuming to handle in the next phase. They should have standard output formats or formats that are easy to convert to these; questions of ethics and rights should have been cleared, and consent forms which clearly indicate future corpus use (see the section hereafter) should be distributed and signed. Zourou (2013) provided a good example of obstacles which may be encountered when collecting data stemming from informal learning situations (such as: Who owns user data in these communities? How accessible is user data? What are the consequences of data ownership and accessibility for research purposes?

Data organization

In this section, we present one way to transform raw data into research data, how to organize themand how to document them in an exhaustive yet informative manner. Besides folders of data coming from the above-mentioned learning design and the research protocol, we detail those gathering participants' productions, ethics and rights information, and the overall organization of the corpus (entitled a global corpus). Later, another corpus type is presented (distinguished corpora), which can be derived from the global corpus after research and analyses have been performed.

Course instantiation

The pedagogical scenario could be perceived as a kind of model of a course, an "abstract class," as phrased in object-oriented languages. When the course takes place, participants (individuals, groups) bring to life this model, i.e., it becomes an "instantiation" of the class. Of course, during this "live" course, events may differ from what was originally planned.

The instantiation component is at the heart of the corpus as this folder regroups all of the data elicitation (Mackey &Gass, 2005). These data are derived from the learning context: all of the participants' productions, including the interaction tracks recorded during the online course. For the Copéas course, this folder includes screen capture videos of the online sessions in Lycéum and the students' reflective reports about the course.

Before regrouping the produced data, a preliminary treatment phase is necessary. Firstly, each resource receives a unique identification code (ID) so that later, in the corpus structuration phase (see hereafter), they can easily be listed and described. A strategic policy is to define IDs which uniquely identify a resource among a set of corpora, e.g., a participant ID may contain the name of the student group to which s/he belongs, the corpus name and course session name—if it is a recording, its mode (video, speech, etc.).

Secondly, all produced data are anonymised through a systematic process. In the Copéas corpus, full names of participants were replaced by participant codes. It is preferable to create meaningful codes which will facilitate data investigation later on. A code can refer to such an aspect as the role of the participant in the course (tutor, student, and researcher), his/her gender, or his/her group ID. One should provide a table that regroups the code, sociolinguistic information, language biography (foreign languages spoken, language level, number of years spent studying the language and context of study) for every participant. Anonymisation also includes modifying any information in the produced data that could lead to the identification of a participant or skew a researcher's analysis of the data. While it is important to anonymise the data, researchers should replace it with meaningful information. It is useful to include the reasons for anonymisation so as to allow interpretations of the interaction. For example, a participant's phone number in a text chat message could be replaced with a code and labelled to highlight that the original information corresponded to a phone number.

Lastly, for the sake of medium and long-term reusability, data collected will be converted into formats independent of their original platform, when the original formats were not open. Several international research associations, including CINES (2014) and Jisc (formerly the Joint Information Systems Committee), involved in the curation and archiving of research data provide clear information about such formats.

Expectations are even greater in regards to participants' interactions that are in text mode, either originally because they have been typed by participants or as the result of transcriptions of speech, for example. Their format will be machine-readable, even structuredin order to detail information about an utterance or a message and relate it to the properties of the environment that integrates this modality. For example, when an LMS includes a discussion forum, every message carries information, such as the author's ID, date of posting, title, message contents, thread, forum name, etc. Rationales for these expectations are related, firstly, to research analysis.

Ethics and rights for OpenData

Releasing a corpus as OpenData means allowing other people the possibility of free use, reuse and distribution. In other words, the user may extract part of the researcher's data, mix this part with data from other sources, add her/his own work to build upon the whole data set and distribute the entire result. Therefore, OpenData relies on two sorts of rights—those related to the data collection and those related to the data release. In other words, data collected need to be free of rights, and, secondly, the corpus creator should give the right to use the corpus to the end-user, thanks to a license that imposes minimal constraints. Indeed, internationally, it is even recommended to avoid putting a licence that forbids commercial use and to waive intellectual property rights (IPR) (Open Knowledge, 2013). Waiving IPRs does not imply that the creators will not be citedor acknowledged. The full bibliographic reference of their work will become prominent in the corpora repository and will guarantee, in the academic world, that end-user researchers can clearly refer to the original creators when submitting their new analysis to a peer-review process.

Collecting data that are free of rights implies that the compiler him/herself has the right to use the resources included in the corpus and that participants waive their rights on what they have produced. Their agreements are obtained once they have individually signed a consent form, distributed after an "enlightenment" procedure (see Mackey &Gass, 2005). During this procedure, researchers have an open discussion with participants, where they explain drawbacks and benefits that may be expected from the course and the research process. For example, for research purposes on gestures, participants can give permission to be directly video-recorded without any post-process blurring. They will also be aware that if they change their minds, they can, at any time, ask for data that concerns them to be removed from the corpus.

The LETEC component that concerns Ethics and Rights contains two distinct parts. The private subfolder regroups all of the informed consent forms signed by the course participants, with contact information. This set of data is not included in the final version of the corpus but rather, due to its confidential nature, is conserved by the corpus compiler. In the second part, the corpus compiler includes a blank example of the informed consent form signed by course participants, besides the corpus licence that details the conditions under which the corpus may be distributed (such as Creative Commons [2015] licences).

Organization of the global corpus

Once the four corpus folders (instantiation, research protocol, learning design, ethics and rights, see LETEC components Figure 10.1) have been organized, with preliminary treatment phases accomplished on the data where necessary, a general document is created. It contains descriptions of each corpus part and crosslinks pieces of information between the different parts (e.g., between the interaction data, research protocol and learning design). It also provides a full index of the resources collected. Each resource is listed, using the previously introduced resource IDS, and a summary of the resource's contents is given. This will help corpus end users determine what data to use, with relation to their specific research question(s).

Lastly, out of the global description, a short corpus description will be extracted so as to provide metadata in formats that website harvesters can recognize and save. The Mulce repository (2013) chose the format created by OLAC (Open Language Archives Community). It is compatible with European CLARIN standards for metadata. This means that metadata concerning all LETEC corpora, including bibliographic citations, appear in these international linguistic resource banks and can be searched for by Internet users.

Post research data and component

Post research often concerns transcriptions of multimodal interactions, in ways which will be presented below. These transcriptions produce a new set of data which will be assembled into a new LETEC, of a distinct type called a distinguished corpus (Reffay et al., 2012). Its size is much reduced, and corresponds to data assembled and produced by a researcher when s/he focuses on a specific research question and aims to publish an article on the specific topic.

A distinguished corpus includes a particular transformation of a selected part of the global corpus—for example, the transformation of a video file into an XML/text file of the transcribed interaction data and its associated metadata. Following transcription, data analyses can be performed. Data from the global corpus are not copied, but instead referred to, and the newly distinguished corpus only adds the transformed data for the specific analysis.

Distinguished corpora help sustain CALL research by valuing the analyses performed by the researcher. The data used for analysis can be presented in parallel with the analysis' results, and distinguished corpora can be cited and referenced in conference papers or published articles. Readers of a researcher's analysis can examine, or reuse the data analysis performed, whilst reading the report of the results.

Corpus publication

Once the content packaging of the corpus is finished, the compiler deposits the corpus in a repository that adheres to the requirements discussed in section 2. This server will provide to the user open access to the information about each corpus stored in the repository with search facilities. It will be connected to harvesters so that its bank of metadata is searchable through each different harvester. It may also offer services such as permalinks to each corpus and data subset, which will identify them in a unique and permanent way.

The Mulce repository (2013) gives access to fifty LETEC corpora coming from more than ten different online learning situations that took place between 2001 and 2013. In May 2012, its size was the following: more than one million tokens, coming from 12000 audio turns, 17000 text chat turns, 3000 blogs, 2000 emails, 2700 discussion forum messages, plus more than 9000 non-verbal acts. The Mulce repository also gives access to more than 200 videos of online interaction sessions. These interactions were produced in a variety of environments (LMS, audio graphic systems, 3D environments, etc.), by groups of learners from different countries, following a range of different pedagogical scenarios. A step-by-step tour of the repository is provided in the article entitled "Discovering LETEC corpora" on the Mulce documentation (2015) website. Needless to say, Mulce encourages other CALL researchers to deposit their corpora in the repository, provided they meet the general criteria outlined here, even if they do not exactly follow certain technical details to which the authors have alluded. Help and discussion will be offered to the depositor.

LETEC contributions to CALL research

The purpose of this section is to present how research on LCI may benefit from the existence of open access corpora. Research is a circular process. For example, LETEC corpora in the Mulce repository have been built out of online learning situations, starting more than thirteen years ago. Data have been reused several times and will be mixed into projects, independent of Mulce, as discussed previously.

Let us start with one of the very first steps in examining online multimodal interactions (see also Chapter 9, this volume), i.e., transcriptions of components of the instantiation part of a LETEC.

Separating transcriptions and analysis steps

Multimodal transcription is a topic discussed across disciplines, for example, in Flewitt et al. (2009). In this article, the authors cite Baldry and Thibault who suggest that "multimodal transcriptions are ultimately based on the assumption that a transcription will help us understand the relationship between a specific instance of a genre, for example a text, and the genre's typical features" (as cited in Flewitt et al., 2009, p. 45). A straightforward interpretation of this statement may induce the idea that all approaches to multimodality should produce their own specific methodological approach to transcriptions. Indeed, the article illustrates various transcription methodologies, from several researchers, that adhere to distinct models of multimodality; in an adhoc fashion, parts of texts, images, photos and hand-made pictures are intertwined in formats such as spreadsheets and word-processing documents, forbidding any kind of comparison or mixing of data. This interpretation of transcription confuses two steps in the research process—the actual transcription and the analysis.

Researchers involved in national or international consortiums on speech and/or multimodal corpora have special interest groups around interoperability (e.g.,Huma-Num, 2014). The idea is that if one understands research as a cumulative process, idiosyncratic models need be compared in order to enhance understanding of human interactions. This implies separating the transcription from analysis processes and using a variety of analysis tools with compatible output formats.

Figure 10.2.belowillustrates this point. It displays a window from the transcription software ELAN (Sloetjes&Wittenburg, 2008) that integrates the video-screen capture of a Copéas session (red box). In this extract, three learners are working in a sub-group to complete a quiz provided by the tutor at the beginning of the session. The tutor comes into the virtual room while one learner is writing an ESL definition using the word processor. Several modes or modalities are being used: audio, text chat (label [3] in the red box) and the word processor (1), plus the iconic system (2), which lists the participants, their status, indicates who is talking and allows simple communication (agreement, disagreement, raise hand, applause, etc.). The transcription process appears in the green box. According to the transcription code used (see Wigham & Chanier, 2015), the researcher defined one layer per participant and per modality (5), i.e., all Learner 1's text chat turns are assembled on the same line, all Learner 1's audio turns on another line, and this is the same for the transcription of his/her actions in the word processor. Transcription is aligned with the video's time, and buttons in (4) provide different ways of selecting parts of the video and of moving between transcriptions layers. Once the transcription is completed, its contents are saved using a text-structured XML format that offers the possibility of later compiling it with transcriptions of other sessions from the same course and/or reusing the file with analysis software.

ELAN is a good example of open-access software. This asset, plus the interoperability one, allows any user, once the distinguished LETEC corpus has been downloaded, to rework on the transcription and add another layer, for example. It is largely used in the aforementioned community on multimodal corpora.

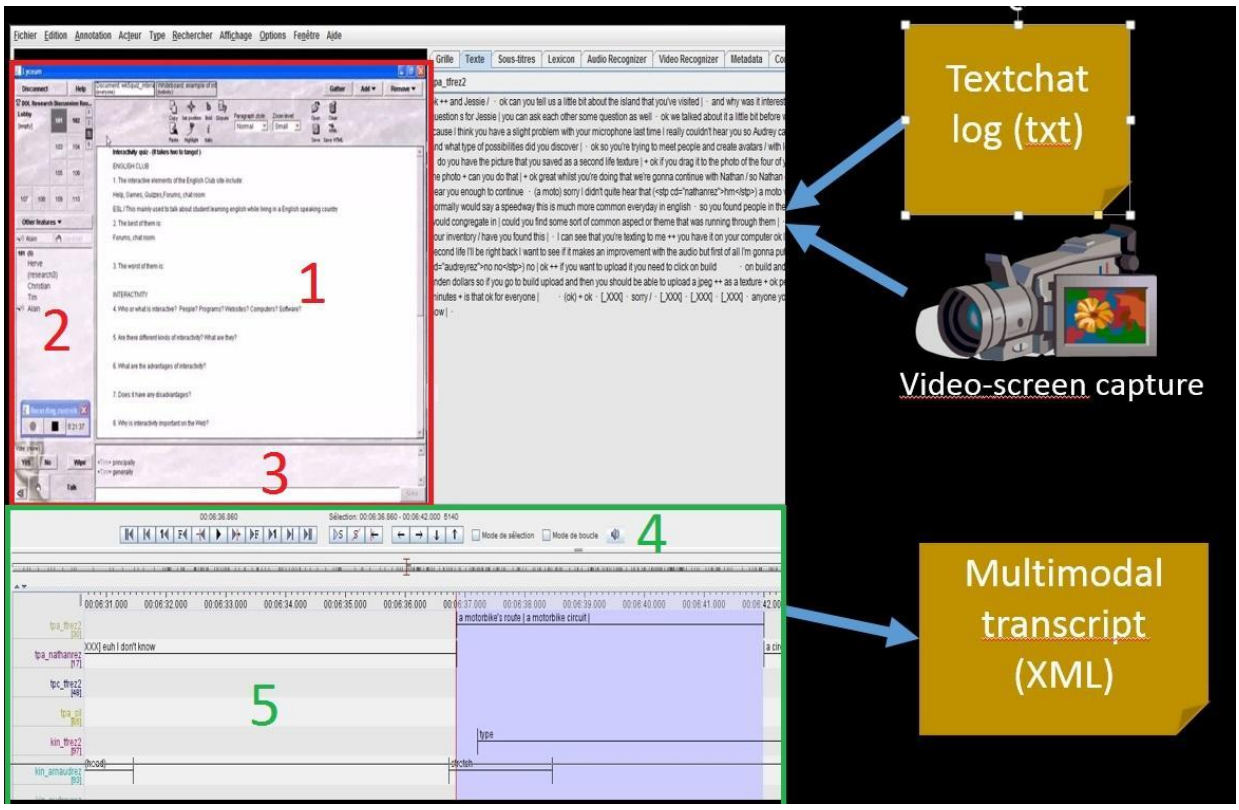

Figure 10.2. Transcript of a Copéas session through the software ELAN, with input and output files.

There is an even subtler methodological question where transcription is concerned: Are online interactions so complex that it is impossible to compare and make adjustments between transcription codes? Let us take an example and consider the code defined when transcribing online learning sessions in 3D environments where participants interact using avatars (Wigham & Chanier, 2015). Shih (2014) provided another approach to the same topic. Are these legitimate differences? Possibly, because it is a new area of research in CALL, where researchers have recourse to a variety of nonverbal communication frameworks. However, if CALL research aims to become more systematic in this area, then the situation may evolve in a manner similar to the area of speech corpora. Whereas textbooks in Second Language Acquisition or Discourse Analysis (e.g.,Schiffrin, 1994) still give the impression that idiosyncratic codes for speech transcription is a normal methodological approach, a community of linguists specialized in speech corpora has developed a common way of transcribing speech (e.g., the CHAT format used in the open access CHILDES repository, MacWhinney, 2009) and has even

included it in a more general framework designed for different text genres called TEI (Text Encoding Initiative [TEI, 2015]). With this extension of XML, a researcher who focuses on a new oral feature may code a new phenomenon whilst being compatible with the rest of the original coding scheme.

Analysis tools and conditions for scientific discussions

Resuming our Copéas example, let us now consider its analysis. Some of the questions the research team had in mind were: Do participants get lost among the multiple possibilities offered by this type of multimodal learning environment? Do they make consistent individual choices? Can they also make collective choices? In the particular sequence alluded to in Figure 2, the workload is distributed among the three learners: one learner types in the shared word processor in order to answer the quiz, and the two others help him orally. Whilst they hesitate on the spelling of a word, the tutor came into the room and typed his corrections into the text chat. This went unnoticed by the learners, and, in turn, the tutor leaves the room. Ciekanski and Chanier (2008) have explained the notion of context which is dynamically built by participants. Relying on this notion developed by Goodwin and Durranti in 1992, their analysis explained that the tutor had been out-of-context. Interestingly, Lamy (2012) imagined the same kind of situation, without referring to any precise data:

Imagine that the tutor led his tutorial via postings in the text-chat while students talked about other topics in the audio channel. It is unlikely that the group would accept such a position for the tutor, and we draw from multimodal social semiotics to help explain why. (p.12)

Discussing alternative explanations with different theoretical references is a very important issue in research, provided that it is supported by data and analysis tools. Figure 3 illustrates our analysis with the open-access tool TATIANA (Dykeet al., 2011), for analysing online interaction from a Computer-Supported Collaborative Learning (CSCL) perspective. To the left of the red line, one can see the same video (top left) with the transcription (bottom left), simply converted from the ELAN-XML output to the TATIANA-XML version. On the right-hand side of Figure 3 appears another view of the desktop, with a view of the modalities used by each participant: one line per participant, one colour per modality (text chat, audio chat, word processor, etc.). This display helps visualize that participants may be out of context, that learners used the word processor in combination with other modalities, which highlights the strategic use of certain modes to facilitate the writing process. The learners also made consistent individual choices to participate in multimodal discourseand to make collective choices. Of course, this analysis has been achieved by examining the whole session, not only theaforementioned extract. The comparison with other sessions and several tools has been explained in Ciekanski and Chanier (2008). The analysis was possible because the output of a first transcription tool became input for a second analysis tool.

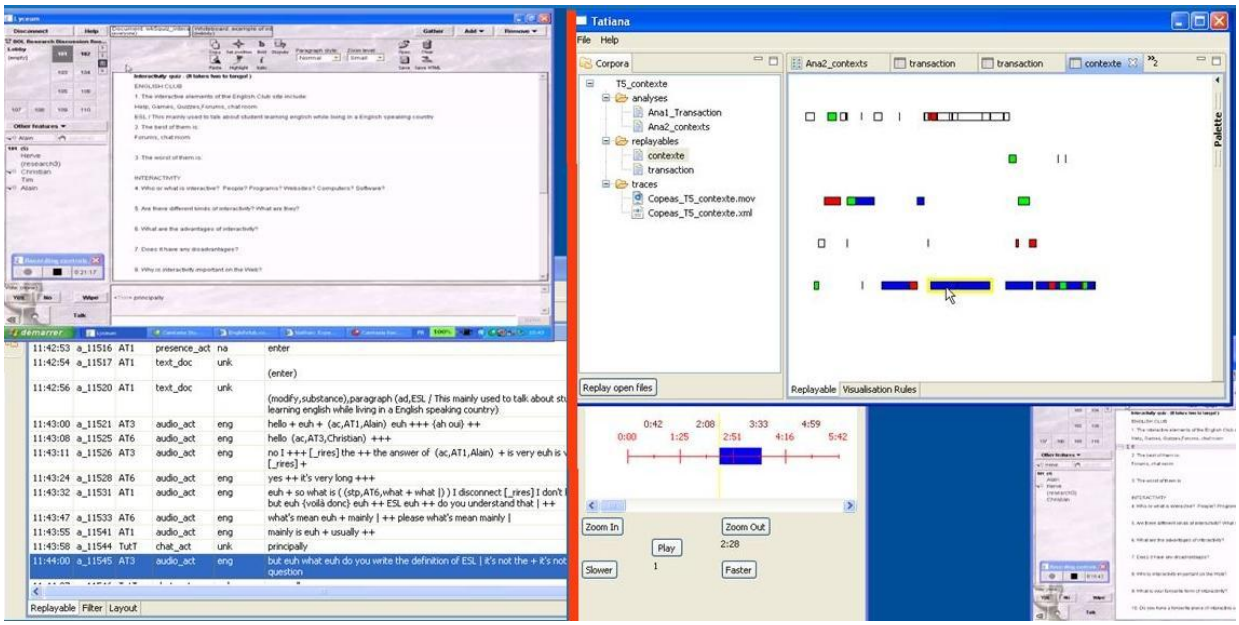

Figure 10.3. Being in and out of context in a multimodal environment. Follow up of example 2 analysed, thanks to the TATIANA software.

How opposite conclusions could be compared

Sindoni (2013) also studied participants' uses of modalities in online environments that integrate audio, video and text chat. She focused on what she termed "mode-switching" when a participant moves from speech to writing or the other way round. She collected dozens of hours of video-screen online conversations that occurred in informal settings (hence not connected to a learning situation). When analysing transcriptions, she observed that participants could be classified according to their preferred interaction mode (oral or written). She also observed that "As anticipated, both speakers and writers, generally carry the interaction forward without mode-switching. This was observed in the whole video corpus" (Sindoni, 2013, section 2.3.5). Hence, she concluded, "those who talked did not write, and those who write did not talk. Turn-taking adheres to each mode" (Sindoni, 2013, section 2.3.5).

In analyses of the Copéas corpus, learners had a preferred mode of expression (oral or written), at least when they were of a beginner level. In contrast with Sindoni (2013), analyses of audio graphic and 3D environments show that learners were mode-switchers (even modality-switchers). Choices of mode/modality depended on the nature of the task that had to be achievedand the tutor's behaviour (e.g.,Wigham& Chanier, 2015).

At this stage, one may expect that scientific discussions could take place between researchers studying online interactions, to debate contradictions, fine differentiations of situations, tasks, etc. In order to allow this, data from the different approaches need to be accessible in standard formats, with publications clearly relating to data and data analyses, and explicit information given about the format of the transcriptions, their

codes and transcription alignments with video. However, Sindoni's (2013) data are not available. The inability to contrast data with other examples, available in open-access formats, is still holding back the scientific advancement of the CALL field.

Coming back to the topic of analysis tools, a researcher who has collected and structured her/his data now has at her/his disposal a wide, rapidly evolvingrange of tools for lexical processing, morpho-syntactic tagging, statistics, discourse analysis, etc. Should the researcher choose open-access toolswith interoperable formats, s/he not only paves the way for circular, multi-analysis research processesbut also contributes to the development of these tools; the teams of researchers who developed them are keen to improve them when confronted with requests based on actual data and analysis attempts. This interface between data-collection and analysis tools is at the heart of what Gray calls "e-science" (cited in Reffay et al., 2011, p. 12) and represents a priority in many different disciplines within the Humanities.

LETEC contributions beyond research in CALL: CMC training for language teachers and linguistics

The need for pedagogical corpora

Extracts of LETEC are currently being developed into resources to train language teachers in how to use CMC tools in their teaching practices. Training teachers out of authentic situations, built upon multimodal materials, is not simply a concern of the language-learning field. Wigham and Chanier (2014) have detailed the extensive experience of the use of classroom video footage in teacher preparation and professional development in face-to-face contexts coming from the fields of physical education, educational sciences, and mathematics, and described the production of several classroom footage video libraries. In the video libraries cited, the resources include two different types of data: (a) raw materials collected during the learning situation (curricular, student work, course planning, instruction and assessment resources),and (b) other *records of practice*(Hatch & Grossman, 2009). These resources include post-course interviews with teachers and also, for example, observation notes made by researchers or trainee-teacher mentors during the class that was filmed. The aim is to give video viewers a sense of what the video footage may fail to capture or details that may have been obscured.

Whilst in other fields, importance is given in teacher training to combining raw materials from experienced teachers' classrooms with research materials, within CALL, CALL-based teacher education is most often delivered through confrontation with research findings and action research (Guichon & Hauck, 2011).

In the first approach, when trainers want students to gain skills in developing online learning situations based on interactive, multimodal environments, they have recourse to the reading of CALL literature disconnected from actual data. Pre-service teachers will not necessarily take the time to question the findings, taking research conclusions as a given. Indeed, the development of an analytic approach to the reading of research

literature takes time, and during training courses, educators do not necessarily have enough time for this process to mature.

The second approach focuses on action research with pre-service teachers participating in experiments and adopting either the role of learners or tutors. Here there is either the assumption that trainees will naturally understand what they need to do or, if greater guidance is given, reflective feedback sessions are often conducted with the trainees. In the latter case, attempts to use the same methodology for both data collection and training purposes are often difficult to manage; trainers face the issue that student materials are often heterogeneous and quickly extracted from the on-going experiment, and pre-service teachers may only consider his/her individual practice.

In the CALL field, training pre-service teachers in CMC out of online learning situations, built upon multimodal materials (carefully analysed with respect to theoretical viewpoints), alongside other records of practice/research data and findings, would be very helpful. Therefore, from the notion of LETEC, which are purely used for research investigations, arose the notion of pedagogical corpora.

An example of pedagogical corpus

Each pedagogical corpus includes a selection of materials from a LETEC corpus and a series of structured teacher-training tasks that have been developed from these materials, based on leads that had been identified in research papers for which the analyses utilized the same data. To illustrate this concept, let us look at a pedagogical corpus, entitled *reflective teaching journals* that was developed from the research Copéas corpus (Wigham & Chanier, 2013).

From the course data and research articles about the project, the need of encouraging trainee-teachers to foster reflective practice through the writing of teaching journals was identified. Journal writing is a prerequisite for developing reflective practice, but it is not a sufficient condition. It only offers a one-sided view of the course situation. A more objective standpoint may come from confronting the journal with other perspectives. In order to make pre-service teachers aware of this situation, the pedagogical corpus focuses on tutors' and students' differing views of successful or unsuccessful collaboration and different perceptions of their online course. The objectives of the corpus are for trainee-teachers to do the following:

- Identify language tutors' and students' differing views of successful online collaboration;
- Summarize the characteristics of successful collaboration and produce a list of implications for practice;
- Appraise the advantages of keeping teaching journals; and
- Compare and contrast reflections from a teaching journal with naturally occurring data (interaction tracks) and researcher-provoked data (student feedback) to analyse whether teachers should base reflections about teaching practice solely on journal entries and personal reactions.

In the pedagogical corpus, the corpus users are guided through a series of reflective activities based on personal experience, extracts from the LETEC: interaction data (audio and video-based), learner questionnaires and both learner and tutor post-course interviews. The online corpus gives the instructions for all tasks, the timing guidelines and suggested student groupings. All tasks can be completed either online or in a teacher training classroom. Figure 10.4 shows a sample task in which users identify characteristics of successful collaboration through the tutor's discourse, using extracts of the reflective journal the tutor kept throughout the Copéas course and an extract of the audio post-course tutor interview.

Activity 3.1

First of all, consult the following resources (rtjounrals-int-TutT-ext1-mp4, rtjounrals-int-TutT-ext2-mp4) that present the tutor's impressions of whether the activities he proposed were collaborative or not. In your notebook, take notes about the characteristics of successful collaboration the tutor gives. Remember that any points he gives about unsuccessful collaboration can be turned on their head to provide pointers for successful collaboration. What reasons does the tutor give for them? Note any examples he gives to illustrate the characteristics you have identified. Do any of the characteristics match those you listed in activity 2?

Resources:

- rtjournals-diary-TutT-pdf This is the tutor's journal that he kept throughout the Copéas course and in which he reflects about tutoring the course online. The journal is in English.
- rtjournals-int-TutT-ext1-mp4 This is a mp4 video of an extract of the audio post-course tutor interview with slides to guide the viewer. A researcher in French conducted the audio interview. The slides are in English. The video lasts 10 minutes 30 seconds.

Figure 10.4. Sample task from a pedagogical corpus (Wigham & Chanier, 2013).

Such pedagogical corpora offer a kind of expert viewpoint (but an expert viewpoint based on research analysis, i.e., coming from a scientific research cycle). Practice in teacher training, coming from the aforementioned fields, shows that it is not enough. Students need to bring their own data (extracts of live sessions and reflective writing) in order to confront these with expert views and other views from classmates as well, the whole process being integrated into a discussion framework, whether online (Barab, Klig&Gray, 2004) or face-to-face. Furthermore, it cannot be a one-shot process but a progressive one. Becoming a teacher implies moving from a peripheral participation to a more centred one, and this process this process must be recognized as legitimate by the community(see Lave & Wenger, 1991). Of course, the teacher training period will not suffice, but the idea is to involve students in a rich process during which they confront expert and novice viewpoints.

Currently, two pedagogical corpora have been developed from two different global LETEC corpora. They can be downloaded from the Mulce repository. They have not yet been used to train teachers. For another approach to using corpora in teacher training, see Chapter 8, this volume.

From learner to general user computer interactions

In this chapter, several references have been made to works and methodologies adopted in linguistics, or corpus linguistics, which influenced CALL research on data. Is this a one-way flow? Does CALL have something to say that could benefit the linguistics field in general? A first refinement of the question could be: Do the language, discourse and texts produced by participants (learners, teachers, etc.) bear similar features (apart from the obvious differences due to the development of the learners' interlanguage, their errors) to those studied in general by linguists interested in computer-mediated discourse?

In order to answer the question, let us consider one type of environment, for example text chat. In the field of linguistics, descriptions of texts and language exist in prototypical works, such as Crystal (2004) in the chapter "The Language of Chatgroups" and its section on synchronous groups. This study aims to give a very general overview of what is actually "the Language of the Internet" as reflected by the book's title. However, when considering text chat coming from CALL, the contents of the turns are strikingly different on both lexical and syntax levels (lexical diversity, use of emoticons or other interaction terms, structures of clauses, of utterances, turn lengths, etc.). The discourse organization is also very different. Whereas nicknames play an important role in informal text chats where users constantly change their nicknames in accordance with their current activities, moods etc., this phenomenon rarely occurs in learning situations. Turns and their combinations (exchanges, transactions, etc.) are managed and structured in a very different manner. In order to support language production in an L2, turn-taking conventions are often adopted[2].

Considering another mode would bring us to the same conclusions. For example, when skimming through corpora where speech is used, either in bimodal environments (text and audio chats) or in richer environments (audio graphic conferencing systems, 3D environments), discrepancies with informal L1 online conversations can be noted concerning a variety of features. To take one but example, speech overlaps in turn taking are not frequent in learning situations. Rationales explaining these differences in the different modes are quite obvious; language teachers organize scenarios beforehand, and tutors interact in ways that support language learners' productions, helping them take risks in a new language while simultaneously alleviating other tasks. CALL research has also begun to show that the orchestration and use of modes and modalities are different to non-educational situations, as previously exemplified in the discussion of Sindoni's work. To some extent, it could be said that multimodality can be "decomposed" to allow some specific modes and modalities to be used in order to focus on specific tasks (for an example, see the focus on writing in Ciekanski &Chanier [2008]). To sum up, the CALL experience of online interactions, supported by its specific corpora, can be of general interest to the whole linguistic community.

---

[2]The reader interested in comparing such differences could access, for example, an informal textchat corpus from Germany (Dortmund Chat Corpus, 2003–2009) or a CALL text chat corpus (Yun & Chanier, 2014).

A common model of CMC interactions

Common interests between CALL and corpus linguistics also concern more abstract levels, such as models of online interaction. Following lessons learnt from the Mulce project (Reffay et al., 2012), researchers are now collaborating with corpus linguists. At a national level, the CoMeRe project (Chanier et al., 2014; CoMeRe, 2015) has brought together corpus linguistics and CALL researchers. The acronym (in French) stands for network-mediated communication, an extension of CMC, in order to include communication through phones, networks and devices. The CoMeRe project has built a kernel corpus in French that represents a variety of network interactions. Several LETEC corpora have been included and structured in the same model alongside corpora of SMS, tweets, Wikipedia discussions, blogs and text chat interactions. The whole set of corpora are released in an open access format.

The CoMeRe team is also working with European researchers specialized in CMC to develop the Interaction Space model (TEI-CMC, 2015) through which to structure these interactions. Briefly, an Interaction Space is an abstract concept, located in time (with a beginning and ending date with absolute time, hence a time frame), where interactions between a set of participants occur within an online location. The online location is defined by the properties of the set of environments used by the set of participants (e.g.,Chanier et al., 2014). Thanks to this model, corpora both from learning and non-learning contexts can, on the one hand, use the same set features to describe the structure and properties of the environment where interactions occurred, the participants (individual, groups), the method for collecting data, for measuring time and durations, etc. On the other hand, in the body of the corpus, the interactions are listed in formats corresponding to their modes (written, oral, or non-verbal). The model is designed by a European group which aims to extend the text model of the Text Encoding Initiative (TEI, 2015) (currently very rich as it encompasses types such as manuscripts, theatre, literature, poems, speech, film and video scripts, etc.) in order to integrate CMC.

Conclusion

When studying LCI in ecological contexts, there are a number of variables that cannot be controlled. These variables make the comparison of scientific results difficult and the replication of a given learning and teaching experience near impossible. This chapter proposed one possible staged methodology to structure raw data from LCI situations into corpora so as to render them comparable, re-analysable and available to the whole research community. The case-study approach adopted allowed us to present the constitution and diffusion of LEarning and TEaching (LETEC) Corpora, using the example of the online Copéas course. In this presentation, we examined the ethical implications of producing corpora as OpenData and suggested ways in which the transcription of LCI and their analysis can become more systematic and comparable.

The LETEC methodology is one methodological proposition to help the CALL field better meet the principles of scientific validity and reliability that are fundamental

cornerstones of the scientific method, yet difficult to achieve in ecological learning situations. More systematic organization of data and its processing is often perceived as time-consuming. However, it requires a mind-set shift whereby individual researchers do not think of producing one-off analyses on individual learning situations but instead look towards long-term team research projects in which corpora, rather than data, are re-used for new analyses, produced from different perspectives, and are reconsidered and cross-referenced from one LCI experiment to another. This would encourage, firstly, a more circular and multi-analysis research approach within the field and, secondly, scientific debate, both within CALL but also more largely within corpus linguistics, which is based on the possibility to reanalyse, verify and extend original findings and to contrast data with other examples from other research teams and different online environments.

References

Barab, S. A., Kling, R., &Gray, J. H. (Eds.). (2004). *Designing for virtual communities in the service of learning*. Cambridge, United Kingdom: Cambridge University Press.

Boulton, A (2011). Data-driven learning: The perpetual enigma. In S. Gozdz-Roszkowski (Ed.), *Explorations across languages and corpora* (pp. 563-580). Frankfurt, Germany: Peter Lang.

Chanier, T., Poudat, C., Sagot, B., Antoniadis, B., Wigham, C. R., Hriba L.,Seddah, D. (2014). The CoMeRe corpus for French: Structuring and annotating heterogeneous CMC genres. *Journal of Language Technology and Computational Linguistics*, *29*(2), 1-31.

Chanier, T. (2013). *EUROCALL 2013, Survey on CALL in the digital humanities: Considering CALL journals, research data*. Paper presented at EUROCALL 2013, Évora, Portugal

Chanier, T., Reffay, C., Betbeder, M-L., Ciekanski, M., & Lamy, M-N. (2009). LETEC (Learning and Teaching Corpus) Copéas [corpus]. Mulce.org: Clermont Université.

Ciekanski, M., & Chanier, T. (2008). Developing online multimodal verbal communication to enhance the writing process in an audio-graphic conferencing environment. *ReCALL*, *20*(2), 162-182. doi:10.1017/S0958344008000426

CINES (2014). Description of resource formats eligible for archiving. National Computing Center for Higher Education. https://www.cines.fr/en/long-term-preservation/expertises/ formats-expertise/facile/

CoMeRe (2015). Repository of computer-mediated communication (CMC) corpora. [webservice] Ortolang: Nancy. Available at http://hdl.handle.net/11403/comere

Creative Commons (2015). Licences for publishing knowledge and data in open access formats [website]. http://creativecommons.org/

Crystal, D. (2004). *The language of the Internet*. Cambridge, United Kingdom: Cambridge University Press.

DARIAH (2013). Digital research infrastructure for arts and humanities [website]. http://www.dariah.eu/

Dortmunder Chat-Korpus (2003-09). German corpus of informal text chat [website]. http://www.chatkorpus.tu-dortmund.de/

DWDS (2013). Das DigitaleWörterbuch der deutschenSprache [website]. http://www.dwds.de

Dyke, G., Lund, K., Jeong, H., Medina, R., Suthers, D. D., van Aalst, J., . . .Looi, C-K. (2011). Technological affordances for productive multivocality in analysis. In H. Spada, G. Stahl, N. Miyake, N. Law, & K. M. Cheng (Eds.), *Connecting Computer-Supported Collaborative Learning to Policy and Practice: Proceedings of the 9th International Conference on Computer-Supported Collaborative Learning (CSCL 2011)* (Vol. I, pp. 454-461). Hong Kong, China: International Society of the Learning Sciences.

Flewitt, R., Hampel, R., Hauck, M., & Lancaster, L. (2009). What are multimodal data and transcription? In C. Jewitt (Ed.), *The Routledge Handbook of Multimodal Analysis* (pp. 40-53). London, United Kingdom: Routledge.

Granger, S. (2004). Computer learner corpus research: Current status and future prospects. In U. Connor & T. Upton (Eds.), *Applied corpus linguistics*: *A multidimensional perspective* (pp. 123-145). Amsterdam, Netherlands: Rodopi.

Guichon, N., & Hauck, M. (2011). Teacher education research in CALL and CMC: More in demand than ever. *ReCALL*, *23*(3), 187-199. doi:10.1017/S0958344011000139

Hatch, T., & Grossman, P. (2009). Learning to look beyond the boundaries of representation. *Journal of Teacher Education*, *60*(1), 70-85. doi:10.1177/0022487108328533

Huma-Num (2014). French national corpora consortiums in various areas of arts and humanities. http://www.huma-num.fr/service/consortium

IMS Learning (2004). Description of learning scenarios (IMS-LD), and data packaging (IMS-CP). IMS Global Learning Consortium. http://www.imsglobal.org/specifications.html

Francis, W. N., &Kučera, H. (1964). *A standard corpus of present-day edited American English, for use with digital computers*. Providence, Rhode Island: Brown University.

Laks, B. (2010). La linguistique des usages: de l'exemplum au datum. In P. Cappeau, H. Choquet, & F. Valetoupolos (Eds.), *L'exemple et le corpus quel statut? Travaux linguistique du Cerlico, 23* (pp. 13-28). Rennes, France: Presses Universitaire de Rennes. Pp. 13-28.

Lamy, M-N. (2012). Click if you want to speak: Reframing CA for research into multimodal conversations in online learning. *International Journal of Virtual and Personal Learning Environments*, *3*(1), 1-18. doi:10.4018/jvple.2012010101

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, United Kingdom: Cambridge University Press.

Mackey, A., &Gass, S. M. (2005). *Second language research: Methodology and design.* New York, NY: Routledge.

MacWhinney, B. (2009). Manual of the CHAT transcription format used in the CHILDES project. http://repository.cmu.edu/cgi/viewcontent.cgi?article=1181&context=psychology

Mulce documentation (2015). Documentation on Mulce repository and Mulce methodology [website]. http://mulce.org

Mulce repository (2013). Repository of learning and teaching (LETEC) corpora [webservice]. Clermont Université: MULCE.org. http://repository.mulce.org

Open Knowledge (2013). Definition of "open" with respect to knowledge and data. http://opendefinition.org/okd/

O'Keefe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge, United Kingdom: Cambridge University Press.

Reffay, C., Betbeder, M.-L., & Chanier, T. (2012). Multimodal learning and teaching corpora exchange: Lessons learned in 5 years by the Mulce project. *International Journal of Technology Enhanced Learning*, *4*(1), 1-20. doi:10.1504/IJTEL.2012.048310

Schiffrin, D. (1994). *Approaches to discourse*. Malden, MA: Blackwell.

Shih, Y.-C. (2014). Communication strategies in a multimodal virtual communication context. *System*, *42*, 34-47. doi:10.1016/j.system.2013.10.016

Sindoni, M. G. (2013). *Spoken and written discourse in online interactions: A multimodal approach*. New York, NY: Routledge.

Sloetjes, H., &Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 816-820.

TEI (2015). Text Encoding Initiative. [website]. http://www.tei-c.org/

TEI-CMC (2015). Computer-Mediated Communication working group of the TEI consortium [website] http://wiki.tei-c.org/index.php/SIG:Computer-Mediated Communication

The Open University. (2001). *Research methods in education*. Milton Keynes, United Kingdom: The Open University.

Wigham, C. R., & Chanier, T. (2013) Pedagogical corpus: Reflective teaching journals. [corpus] Mulce.org: Clermont Université. [oai: mulce.org:mce-peda-rtjournals; http://repository.mulce.org].

Wigham, C.R. & Chanier, T. (2014). Pedagogical corpora as a means to reuse research data and analyses in teacher-training. In J. Colpaert, A. Aerts, & M. Oberhofer (Eds), *Proceedings of the sixteenth international CALL conference, 7-9 July 2014* (pp. 360-365). Antwerp, Belgium: University of Antwerp.

Wigham, C. R., & Chanier, T. (2015). Interactions between text chat and audio modalities for L2 communication and feedback in the synthetic world Second Life. *Computer AssistedLanguage Learning*, *28*(3), 260-283. doi:10.1080/09588221.2013.851702

Yun, H., & Chanier, T. (2014). Corpus d'apprentissage FAVI (Français académique virtuel international). Banque de corpus CoMeRe.[corpus] Ortolang.fr: Nancy.

Zourou, K. (2013). Research challenges in informal social networked language learning communities. *eLearning Papers*, *34*, 1-11.